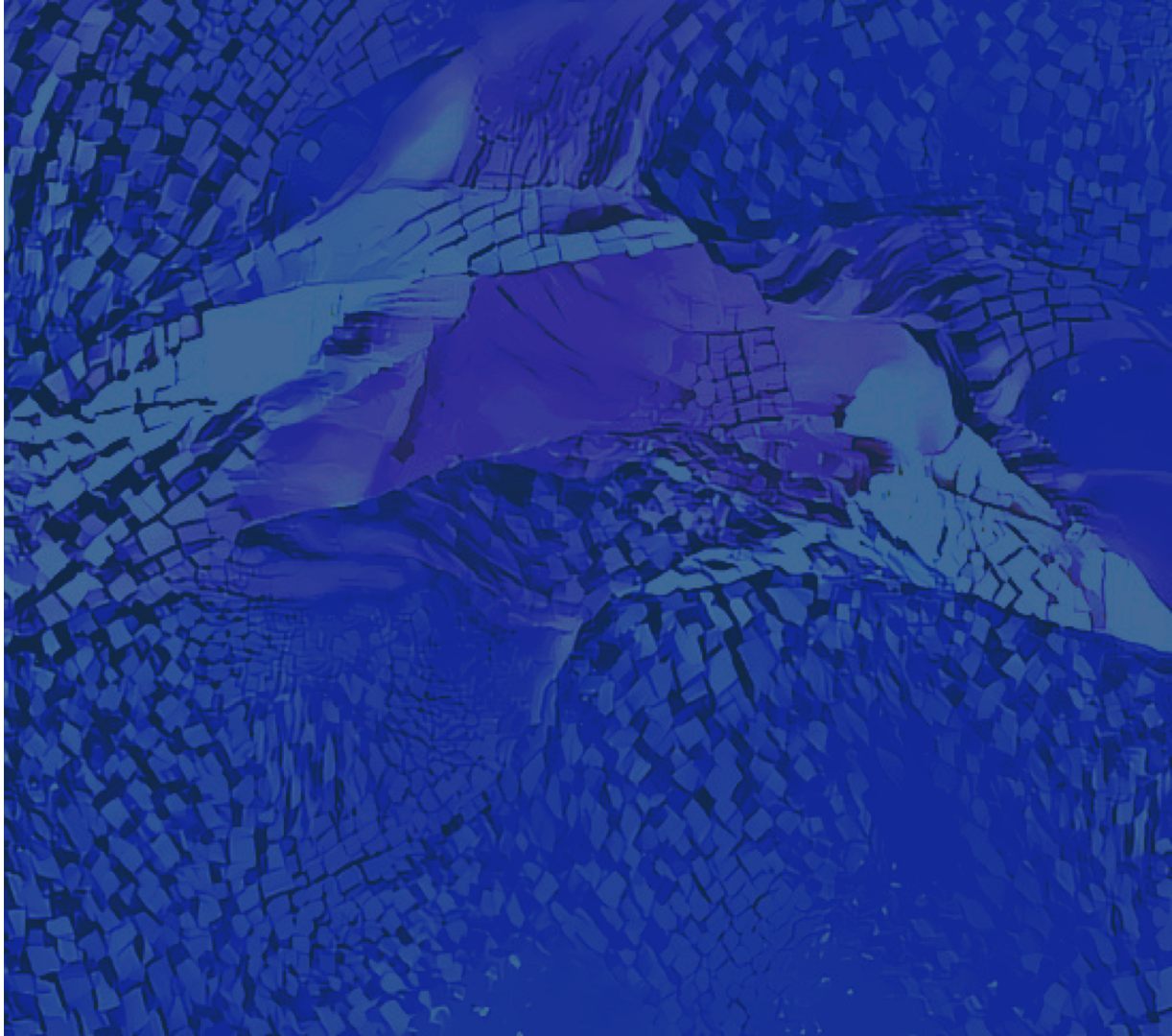
DisSent: Learning Sentence Representations from Explicit Discourse Relations. ACL 2019

Antoine SIMOULIN

23/04/2020

INRIA Working Group



DisSent

Sentence embedding can be used for multiple use cases: paraphrase detection, summarization, knowledge-base population, question-answering, automatic message forwarding, and metaphoric language, to name a few.

However, Existing models either train on a **vast amount of text**, or require costly, manually curated sentence relation datasets.

Discourse markers annotate deep conceptual relations between sentences. Learning to predict them may thus represent a strong training task for sentence meanings.

This task is an interesting **intermediary between two recent trends.**

[1] Allen Nie, Erin Bennett, Noah D. Goodman: DisSent: Learning Sentence Representations from Explicit Discourse Relations. ACL (1) 2019: 4497-4510

Discourse Prediction Task

Discourse Prediction Task to train sentence embeddings

Train a sentence encoding model to learn embeddings for each sentence in a pair such that a classifier can **identify**, based on the embeddings, **which discourse marker was used to link the sentences**

Use the DisSent task to **fine-tune larger pre-trained models such as BERT**.

Data Collection

S1	marker	S2
Her eyes flew up to his face.	and	Suddenly she realized why he looked so different.
The concept is simple.	but	The execution will be incredibly dangerous.
You used to feel pride.	because	You defended innocent people.
Ill tell you about it.	if	You give me your number.
Belter was still hard at work.	when	Drade and barney strolled in.
We plugged bulky headsets into the dashboard.	so	We could hear each other when we spoke into the microphones.
It was mere minutes or hours.	before	He finally fell into unconsciousness.
And then the cloudy darkness lifted.	though	The lifeboat did not slow down.

Table 3: **Example pairs** from our Books 8 dataset.

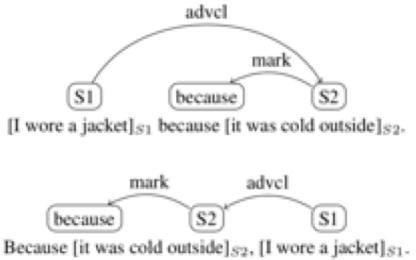


Figure 1: **Dependency patters for extraction:** While the relative order of a discourse marker (e.g. *because*) and its connected sentences is flexible, the dependency relations between these components within the overall sentence remains constant. See Appendix A.1 for dependency patterns for other discourse markers.

Data Collection

Marker	Extracted Pairs	Percent (%)
but	1,028,995	21.86
and	1,020,316	21.68
as	748,886	15.91
when	527,031	11.20
if	472,852	10.05
before	218,305	4.64
because	167,358	3.56
while	161,818	3.44
though	104,218	2.21
after	95,847	2.04
so	76,940	1.63
although	37,511	0.80
then	16,429	0.35
also	16,365	0.35
still	13,421	0.29
Total	4,706,292	100.0

Table 2: Number of pairs of sentences extracted from BookCorpus for each discourse marker and percent of each marker in the resulting dataset.

Task	# of examples	# of tokens
SNLI + MNLI	0.9M	16.3M
DisSent Books 5	3.2M	63.5M
SkipThought	—	800M
BERT MLM/NSP	—	3300M

Table 1: Training data size (in millions) in each pre-training task. DisSent Books 5 only uses 5 discourse markers instead of all.

That markers with similar meanings tended to be confusable with one another.

Related Work – Unsupervised Learning

Supervision given discourse story line

Figure 1. *Skip-thougth vectors: generative objective for next and previous sentence [1] or just BoW prediction [6]*

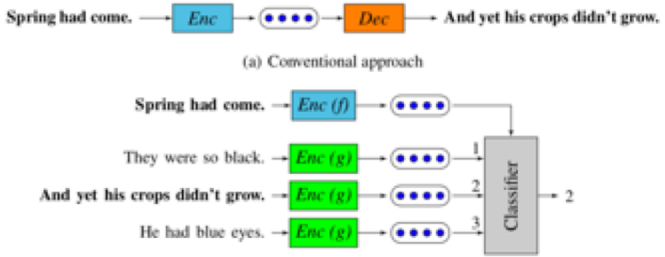


Figure 2. *Quick-thougth vectors: discriminative objective for next and previous sentence given negative samples [5]*

Supervision given altered language model

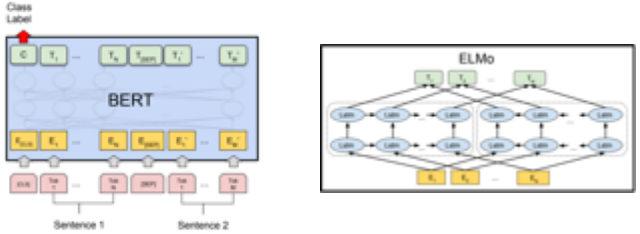


Figure 3. *Mask language models, Bert & Co, ELMo, Open AI GPT [2, 3, 4].*

*[1] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, Sanja Fidler: **Skip-Thought Vectors**. NIPS 2015: 3294-3302*
*[2] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer: **Deep Contextualized Word Representations**. NAACL-HLT 2018: 2227-2237*
*[3] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. **Language models are unsupervised multitask learners**.*
*[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. NAACL-HLT (1) 2019: 4171-4186*
*[5] Lajanugen Logeswaran, Honglak Lee: **An efficient framework for learning sentence representations**. ICLR (Poster) 2018*
*[6] Felix Hill, Kyunghyun Cho, Anna Korhonen: **Learning Distributed Representations of Sentences from Unlabelled Data**. HLT-NAACL 2016: 1367-1377*

Related Work – Supervised Learning

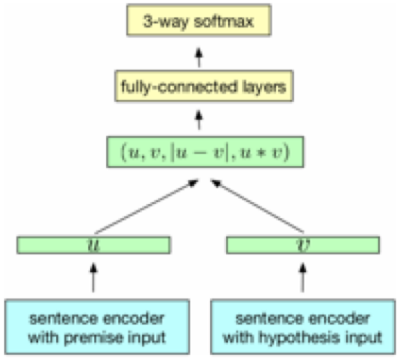


Figure 1. *Siamese architecture For Natural Language Inference task [1, 2]*

Natural language Inference task shows good transfer learning properties. Model are usually trained on the concatenation of SNLI and Multi-Genre NLI (All NLI) with either LSTM or Bert encoder [1, 2]

[1] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault , Antoine Bordes: ***Supervised Learning of Universal Sentence Representations from Natural Language Inference Data****. EMNLP 2017: 670-680*

[2] Nils Reimers, Iryna Gurevych: ***Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks****. EMNLP/IJCNLP (1) 2019: 3980-3990*

Results on SICK dataset

Model	Dim	Hrs	MR	CR	SUBJ	MPQA	TREC	MRPC		SICK-R			SICK-E
								Acc	F1	r	ρ	MSE	
Context sentences prediction													
FastSent	< 500	2	70.8	78.4	88.7	80.6	76.8	72.2	80.3	---	---	---	---
FastSent + AE	< 500	2	71.8	76.7	88.8	81.5	80.4	71.2	79.1	---	---	---	---
Skipthought	4800	336	76.5	80.1	93.6	87.1	92.2	73.0	82.0	85.8	79.2	26.9	---
Skipthought + LN	4800	672	79.4	83.1	93.7	89.3	88.4	---	---	85.8	78.8	27.0	**79.5**
Quickthoughts	4800	11	80.4	**85.2**	93.9	**89.4**	**92.8**	76.9	84.0	**86.8**	**80.1**	**25.6**	---
Sentence relations prediction													
DisSent Books 5	4096	---	80.2	85.4	93.2	90.2	91.2	76.1	---	84.5	---	---	83.5
DisSent Books 8	4096	---	79.8	85.0	93.4	**90.5**	93.0	76.1	---	85.4	---	---	83.8
InferSent	4096	---	81.1	86.3	92.4	90.2	88.2	**76.2**	**83.1**	**88.4**	---	---	**86.3**
Sentence-BERT (base)	768	---	**83.6**	**89.4**	**94.4**	89.9	89.6	76.0	---	72.9	---	---	---
Universal Sentence Encoder	---	---	80.1	85.2	94.0	86.7	**93.2**	70.1	---	76.7	---	---	---
Transformers / Altered language models													
Avg. BERT embeddings	768	---	**78.7**	**86.2**	**94.4**	**88.7**	**92.8**	69.5	---	**78.4**	---	---	---
BERT [CLS]	768	---	**78.7**	84.9	94.2	88.2	91.4	**71.1**	---	42.6	---	---	---
Multi-task learning													
LSTML			**82.5**	**87.7**	**94.0**	**90.9**	**93.0**	**78.6**	---	**88.8**	---	---	**87.8**

Table 1. Results on the SentEval benchmark [1]. **Bold** figures indicate best scores in the category. **Underlined** figures indicate best model overall.

*[1] Alexis Conneau, Douwe Kiela: **SentEval: An Evaluation Toolkit for Universal Sentence Representations**. LREC 2018*

Results - Implicit Discourse Relation Task

Model	IMP	MVU
Sentence Encoder Models		
SkipThought (Kiros et al., 2015)	9.3	57.2
InferSent (Conneau et al., 2017)	39.3	84.5
Patterson and Kehler (2013)	—	86.6
DisSent Books 5	40.7	86.5
DisSent Books 8	41.4	**87.9**
DisSent Books ALL	**42.9**	87.6
Fine-tuned Models		
BERT	52.7	80.5
BERT + MNLI	53.7	80.7
BERT + SNLI + MNLI	51.3	79.8
BERT + DisSent Books 5	**54.7**	81.6
BERT + DisSent Books 8	52.4	80.6
BERT + DisSent Books ALL	53.2	**81.8**
Previous Single Task Models		
Word Vectors (Qin et al., 2017)	36.9	74.8
Lin et al. (2009) + Brown Cluster	40.7	—
Adversarial Net (Qin et al., 2017)	**46.2**	—

Table 7: **Discourse Generalization Tasks using PDTB:** We report test accuracy for sentence embedding and state-of-the-art models.

Are sentence pairs with explicitly marked relations are qualitatively different from those where the relation is left implicit?

Discussion

👍

Method really **clear** and easy to understand.

Code, model and data available:
https://github.com/windweller/DisExtract

Linguistic insight: Idea is rather new, similarity with Rethorical Discourse Theory (RST) but much simpler.

Training **dataset is much smaller** compared to similar training method. The dataset constitution is also semi-automatic

👎

Encoder is a simple LSTM, difficult to **compare with large pre-trained models**

Compare **pair of sentences** but only adjacent one. Is it sufficient to infer relation between non-adjacent sentences?

Only **quantitative analysis on meta-benchmark**, difficult to isolate the reason for performance. No ablation study.

Training time not reported, could be useful for comparison. Should be reasonable though.